

EL691512224US

EL764085881

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION FOR LETTERS PATENT

**A Pull Based, Intelligent Caching System and Method
for Delivering Data Over a Network**

Inventor(s):

Gregory Burns

Paul Leach

ATTORNEY'S DOCKET NO. MS1-095US

Ans A/

1 **TECHNICAL FIELD**

2 This invention relates to network systems, and particularly public network
3 systems, such as the Internet. More particularly, this invention relates to methods
4 which improve distribution of streaming continuous data (e.g., audio and video
5 data) from a content provider over a network to a subscriber's computer or other
6 content rendering unit.

7

8 **BACKGROUND OF THE INVENTION**

9 Public networks, and most notably the Internet, are emerging as a primary
10 conduit for communications, entertainment, and business services. The Internet is
11 a network formed by the cooperative interconnection of computing networks,
12 including local and wide area networks. It interconnects computers from around
13 the world with existing and even incompatible technologies by employing common
14 protocols that smoothly integrate the individual and diverse components.

15 The Internet has recently been popularized by the overwhelming and rapid
16 success of the World Wide Web (WWW or Web). The Web is a graphical user
17 interface to the Internet that facilitates interaction between users and the Internet.
18 The Web links together various topics in a complex, non-sequential web of
19 associations which permit a user to browse from one topic to another, regardless of
20 the presented order of topics. A "Web browser" is an application which executes
21 on the user's computer to navigate the Web. The Web browser allows a user to
22 retrieve and render hypermedia content from the WWW, including text, sound,
23 images, video, and other data.

24 One problem facing the continued growth and acceptance of the Internet
25 concerns dissemination of streaming continuous data, such as video and audio

1 content. Data is delivered and rendered to users in essentially two formats. The
2 first format, referred to as "block data," entails downloading the entire data set to
3 local storage and then rendering the data from the locally stored copy. A second
4 delivery format, known as "streaming data," entails sending bits of data
5 continuously over the network for just-in-time rendering.

6 Computer network users have been conditioned through their experiences
7 with television and CD-ROM multimedia applications to expect instantaneous
8 streaming data on demand. For technical reasons, however, the Internet is often
9 unable to deliver streaming data. This inability is most pronounced for video data.
10 In the Internet context, there is often long delays between the time video content is
11 requested and the time when the video content actually begins playing. It is not
12 uncommon to wait several minutes for a video file to begin playing. In essence,
13 for factors discussed below, video data is traditionally delivered as "block data"
14 over the Internet and thus requires that the entire file be downloaded prior to
15 rendering.

16 The inability to provide streaming data is a result of too little bandwidth in
17 the distribution network. "Bandwidth" is the amount of data that can be moved
18 through a particular network segment at any one time. The Internet is a
19 conglomerate of different technologies with different associated bandwidths.
20 Distribution over the Internet is usually constrained by the segment with the lowest
21 available bandwidth.

22 Fig. 1 shows a model of a public network system 20, such as the Internet.
23 The network system 20 includes a content server 22 (e.g., a Web server) which
24 stores and serves multimedia data over a distribution network 24. The network
25 system 20 also has regional independent service providers (ISPs) or point of

1 presence (POP) operators, as represented by ISP 26, which provide the
2 connectivity to the primary distribution network 24. Many users, as represented
3 by subscriber computers 28, 30, and 32, are connected to the ISP 26 to gain access
4 to the Internet.

5 The ISP 26 is connected to the distribution network 24 with a network
6 connection 34. In this example illustration, the network connection 34 is a "T1"
7 connection. "T1" is a unit of bandwidth having a base throughput speed of
8 approximately 1.5 Mbps (Megabits per second). Another common high bandwidth
9 connection is a T3 connection, which has a base throughput speed of
10 approximately 44.7 Mbps. For purposes of explaining the state of the technology
11 and the practical problems with providing real-time streaming data over the
12 Internet, it is sufficient to understand that there is also a limited bandwidth
13 connection between the content server 22 and the distribution network 24.

14 The subscriber computers 28, 30, and 32 are connected to their host ISP 26
15 via home entry lines, such as telephone or cable lines, and compatible modems.
16 As examples of commercially available technology, subscriber computer 28 is
17 connected to ISP 26 over a 14.4K connection 36 which consists of a standard
18 telephone line and a V.32bis modem to enable a maximum data rate of 14.4 Kbps
19 (Kilobits per second). Subscriber computer 30 is connected to the ISP 26 with a
20 28.8K connection 38 (telephone line and V.34 modem) which supports a data rate
21 of 28.8 Kbps. Subscriber computer 32 is connected to the ISP 26 with an ISDN
22 connection 40 which is a special type of telephone line that facilitates data flow in
23 the range of 128-132 Kbps. Table 1 summarizes connection technologies that are
24 available today.

Table 1: Connection Technologies and Throughput

<u>Connection Type</u>	<u>Base Speed (Kbps)</u>
V.32bis modem	14.4
V.34 modem	28.8
56K Leased Line	56
ISDN BRI (1 channel)	56-64
ISDN BRI (2 channels)	128-132
Frame Relay	56-1,544
Fractional T1	256-1,280
ISDN PRI	1,544
Full T1 (24 channels)	1,544
ADSL	2,000-6,000
Cable Modem	27,000
T3	44,736

With a T1 connection to the primary distribution network 24, the ISP 26 can facilitate a maximum data flow of approximately 1.5 Mbps. This bandwidth is available to serve all of the subscribers of the ISP. When subscriber computer 28 is connected and downloading data files, it requires a 14.4 Kbps slice of the 1.5 Mbps bandwidth. Subscriber computers 30 and 32 consume 28.8 Kbps and 128 Kbps slices, respectively, of the available bandwidth.

The ISP can accommodate simultaneous requests from a number of subscribers. As more subscribers utilize the ISP services, however, there is less available bandwidth to satisfy the subscribers requests. If too many requests are received, the ISP becomes overburdened and may not be able to adequately service

1 the requests in a timely manner, causing frustration to the subscribers. If latency
2 problems persist, the ISP can purchase more bandwidth by adding additional
3 capacity (e.g., upgrading to a T3 connection or adding more T1 connections).
4 Unfortunately, adding more bandwidth may not be economically wise for the ISP.
5 The load placed on the ISP typically fluctuates throughout different times of the
6 day. Adding expensive bandwidth to more readily service short duration high-
7 demand times may not be profitable if the present capacity adequately services the
8 subscriber traffic during most of the day.

9 The latency problems are perhaps the most pronounced when working with
10 video. There are few things more frustrating to a user than trying to download
11 video over the Internet. The problem is that video requires large bandwidth in
12 comparison to text files, graphics, and pictures. Additionally, unlike still images
13 or text files, video is presented as moving images which are played continuously
14 without interruption. Video typically requires a 1.2 Mbps for real-time streaming
15 data. This 1.2 Mbps throughput requirement consumes nearly all of a T1
16 bandwidth (1.5 Mbps). Accordingly, when multiple subscribers are coupled to the
17 ISP and one subscriber requests a video file, there is generally not enough capacity
18 to stream the video in real-time from the content server 22 over the Internet to the
19 requesting subscriber. Instead, the video file is typically delivered in its entirety
20 and only then played on the subscriber computer. Unfortunately, even
21 downloading video files in the block data format is often inconvenient and usually
22 requires an excessive amount of time.

23 Consider the following example. Suppose a subscriber wishes to access the
24 CNN Web site on the Internet for an account of recent news. As part of the news
25 materials, CNN provides a twenty second video clip of an airplane hijacking

1 incident. At 1.2 Mbps, the 20 second video clip involves downloading a 24 Mbyte
2 file over the Internet. If the user has a modest 14.4 Kbps connection, it would take
3 approximately 28 minutes to download the entire file.

4 Now, assume that the subscriber/ISP connection is sufficiently large to
5 handle real-time video streaming of the video file, meaning that the subscriber
6 computer can render the video data as it is received from the ISP. Despite the
7 bandwidth of the subscriber/ISP connection, real-time video streaming may still be
8 unachievable if the T1 connection 34 between the ISP 26 and the distribution
9 network 24 is unable, or unwilling due to policy reasons, to dedicate 1.2 Mbps of
10 its bandwidth to the video file. Requests for the CNN video clip made during peak
11 traffic times at the ISP most certainly could not be accommodated by the
12 ISP/network connection. Since adding more bandwidth may be a poor investment
13 for the ISP, the ISP may have no economic incentive to remedy the latency
14 problem. The result is that some users might be inconvenienced by the lack of
15 ability to receive streaming video despite their own connection to the ISP being
16 capable of accommodating streaming video.

17 The latency problem is further aggravated if the connection between the
18 content server 22 and the distribution network 24 is equally taxed. The lack of
19 sufficient bandwidth at the content server/network link could also prevent real-
20 time video streaming over the Internet, regardless of the bandwidths of the
21 network/ISP link or the ISP/subscriber link. If all links lack sufficient bandwidth,
22 the latency problem can be compounded.

23 One solution to this problem is to provide local cache storage at the ISP. As
24 subscribers request files from the Internet, the ISP caches the files locally so that
25 subsequent requests are handled in a more expeditious manner. This process is

1 known as "on-demand caching." Local on-demand caching methods improve the
2 ability to deliver video content over the Internet. When the first subscriber
3 requests the CNN video clip of the airplane hijacking incident, the ISP requests the
4 video clip from the CNN server, and facilitates delivery of the video clip to the
5 requesting subscriber. The ISP also caches the video clip in its own memory.
6 When any subsequent subscriber requests the same CNN video clip, the ISP serves
7 the local version of the video clip from its own cache, rather than requesting the
8 clip from the CNN server. If the subscriber computer has a high bandwidth
9 connection with the ISP, the locally stored video clip can be served as continuous
10 streaming video data for instantaneous rendering on the subscriber computer.

11 A drawback of the on-demand caching method is that the first requesting
12 subscriber is faced with the same latency problems described above. All
13 subsequent subscribers have the benefit of the cached version. However, if the
14 initial delay is too long, there may not be any subscriber who is willing to assume
15 the responsibility of ordering the video file and then waiting for it to download.

16 Accordingly, there remains a need to develop improved techniques for
17 facilitating distribution of streaming video over public networks, such as the
18 Internet.

19

20 **SUMMARY OF THE INVENTION**

21 This invention provides improved methods for delivering large amounts of
22 data, such as streaming audio and video data, over a network, such as the Internet.
23 According to one aspect, the method involves an intelligent, pre-caching and pre-
24 loading of frequently requested content to the local service provider (e.g., ISP or
25 LAN network server) prior to peak demand times when the content is likely to be

1 requested by the subscribers. In this manner, the frequently requested content is
2 already downloaded and ready to be served to the subscribers before they actually
3 request it. When the content is finally requested, the data is streamed continuously
4 in real-time for just-in-time rendering at the subscriber. This eliminates the
5 latency problems of prior art systems because the subscribers do not have to wait
6 for the downloading of video and audio files over the Internet. Moreover,
7 intelligently pre-caching content before peak demand times is more effective than
8 traditional on-demand caching because the content is available to the first
9 subscriber who requests it.

10 In one implementation, the network system includes a content provider
11 connected to local service providers via a distribution network. The local service
12 providers facilitate delivery of the content from the content provider to multiple
13 subscribers. The local service providers are configured to request certain content
14 from the content provider prior to a peak time when the subscribers are likely to
15 request the content. The content is downloaded from the content provider during
16 non-peak hours and cached at the local service providers for serving to the
17 subscribers during the ensuing peak time.

18 The local service provider includes a processing control unit, a cache
19 memory, and a continuous media server. A hit recording module executes on the
20 processing control unit to record requests for particular content from the
21 subscribers. In the Internet context, these requests are submitted in the form of
22 URLs (universal resource locators) for target resources located on the Web. A
23 pattern recognizer detects behavior patterns based on subscriber requests to
24 determine which content the subscribers are most likely to request and when. A
25 scheduler then schedules requests for the frequently requested content from the

1 content provider at a selected time prior to the peak demand time for that content.
2 These requests are posted to the content provider at their scheduled times, and the
3 content provider downloads the content during the off-hours prior to the peak time.

4 When the content is received from the content provider, the local service
5 provider stores the content in the cache memory. For instance, the content might
6 be a Web page from a frequently visited Web site. Web pages are typically
7 designed as hypermedia documents to provide rich multimedia presentations which
8 blend text, images, sound, and video. If the Web page references or includes
9 continuous data files, such as audio or video files, these files are stored in a
10 continuous media server. The target specifications embedded in the Web page to
11 reference the continuous data files are modified to reference the local copy of the
12 continuous data files, as opposed to the original location of the files at the Web
13 site.

14 During the ensuing peak time, the processing control unit serves the target
15 resources maintained in the cache memory to the subscribers. If any subscriber
16 clicks on or otherwise activates a link to an audio or video file, the requested file is
17 served as a continuous stream of data from the continuous media server at the ISP.
18 In this manner, the continuous video or audio data stream can be rendered just-in-
19 time by the subscriber.

20 Another aspect of this invention involves supplementing the primary
21 Internet connection owned by the ISP with a delivery of content over a secondary
22 network. This supplemental delivery effectively increases bandwidth between the
23 content provider and the local service provider. In the described implementation,
24 the content provider broadcasts additional content over a broadcast satellite
25 network to the local service provider. The broadcast communication link offers

1 additional bandwidth at a fraction of the cost that would be incurred if the local
2 service provider installed additional Internet connections, such as T1 or T3
3 connections. The broadcasted content is stored at the local service provider and
4 served during peak times to afford continuous audio streaming to the subscribers.

5

6 **BRIEF DESCRIPTION OF THE DRAWINGS**

7 Fig. 1 is a diagrammatic illustration of a network system which is used to
8 explain the present state of Internet technology.

9 Fig. 2 is a diagrammatic illustration of a network system constructed
10 according to one implementation of this invention.

11 Fig. 3 is a diagrammatic illustration of a network system constructed
12 according to another implementation of this invention.

13 Fig. 4 is a block diagram of the functional components in a local service
14 provider in the network system.

15 Fig. 5 is a flow diagram of a method for operating the local service
16 provider.

17 Fig. 6 is a diagrammatic illustration of a network system according to still
18 another implementation of this invention.

19 The same reference numbers are used throughout the figures to reference
20 like components and features.

21

22 **DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT**

23 Fig. 2 shows a public network system 50. It includes multiple content
24 servers, as represented by content server 52, which store and serve content over a
25 network 54. The content server 52 serves content in the form of text, audio, video,

1 graphic images, and other multimedia data. In the Internet context, the content
2 servers might represent Web sites which serve or multicast content in the form of
3 hypermedia documents (e.g., Web page) which link text, images, sounds, and
4 actions in a web of associations that permit a user to browse through related
5 topics, regardless of the presented order of the topics. The content server 52 might
6 alternatively represent headend servers for a cable company which transmit video
7 content over a cable network, or an audio server for a radio station that sends
8 audio data over the network. The content server 52 might further represent servers
9 for educational institutions, public agencies, libraries, merchants, or any other
10 public or private organizations which serve or multicast information over the
11 network.

12 The network 54 is a high-speed, high-bandwidth interactive distribution
13 network, and can be representative of the Internet. Traffic over the network 54 is
14 organized according to protocols which define how and when data is moved. One
15 example protocol is the transmission control protocol/Internet protocol (TCP/IP)
16 which forms the backbone of the Internet. The network 54 might be implemented
17 using various physical mediums, including wirebased technologies (e.g., cable,
18 telephone lines, etc.) and wireless technologies (e.g., satellite, cellular, infrared,
19 etc.). The network is operated according to high-speed switching services,
20 including connection-oriented network services (e.g., frame relay, asynchronous
21 transfer mode (ATM), etc.) and connectionless services (e.g., switched
22 multimegabit data service, etc.). These switching services support connection
23 speeds of several Megabits per second (Mbps), up to Gigabits per second (Gbps).
24 At these speeds, the network 54 is capable of supporting streaming video data
25 which requires 1.2 Mbps.

1 Many independent service providers (ISPs), as represented by ISP 56,
2 function as terminal connections or "on-ramps" to the high-speed network 54. The
3 ISP 56 acts as an intermediary between the subscribers 58 and 60 and the network
4 54. The ISP 56 has a network port 62 which provides a high-speed, high-
5 bandwidth connection 64 to the network 54. The ISPs segment and rent portions
6 of the bandwidth to the multiple subscribers 58 and 60 so that the subscribers do
7 not individually need to purchase and maintain their own network connections.
8 The ISPs 56 may also be referred to as point of presence (POP) servers, and the
9 names "ISP" and "POP" are used interchangeably in this disclosure.

10 The subscriber personal computers (PCs) 58 and 60 are individually
11 connected to the ISP 56 by permanent or sessional dial-up connections.
12 Conventional telephone or cable lines and compatible modems are used to form
13 the connections 66, 68. Examples of suitable technologies include HFC, ISDN,
14 POTS, and ADSL. The ISP 56 has network terminal switching equipment 70 to
15 accommodate the connections to the subscriber PCs 58, 60.

16 The ISP 56 also has a cache server 72 and a continuous media server
17 (CMS) 74. The cache server 72 is configured as a conventional database server
18 having processing capabilities, including a CPU (not shown), and storage 78. As
19 one example, the cache server 72 is implemented as a SQL (Structure Query
20 Language) database. The cache server 72 caches Internet resources, such as those
21 requested by subscriber computers 58, 60, that have been downloaded from the
22 content provider 52 to allow localized serving of those resources.

23 The CMS 74 is a server designed particularly for serving continuous data
24 streams, such as video data and audio data, in an ordered and uninterrupted
25 manner. As one example implementation, the continuous media server is

1 configured as a disk array data storage system consisting of many large capacity
2 storage disks with video and audio data streams stored digitally thereon. The
3 locations of the video and audio data streams are kept in a memory map and each
4 video and audio data stream is accessed through pointers to the particular memory
5 location. To serve the audio or video data, the processor 80 grabs the pointer to
6 the video stream and begins retrieving the video from the storage disk 82 and
7 streaming it over the communication line 66, 68 to the requesting subscriber
8 computing unit.

9 Fig. 3 shows a network system 90 which is implemented in a local area
10 network (LAN) configuration. This implementation is exemplary of how a
11 company or multi-user organization might be connected to the Internet. The
12 network system 90 differs from the system of Fig. 2 in that the local service
13 provider which facilitates the on-ramp connection to the high-speed, high-
14 bandwidth network 54 is itself a local server 92 on a LAN 94. The LAN 94 can be
15 constructed using conventional network topologies, such as Ethernet. The LAN
16 network server 92 has a network port 96 which enables a high-speed, high-
17 bandwidth connection 98 to the network 54. The cache server 72 and CMS 74 are
18 connected to the LAN 94. Workstations or other computing units 100, 102 are
19 connected to the LAN 94 and are served by the LAN network server 92 in regards
20 to Internet access. In this configuration, the LAN users of workstations 100, 102
21 have access to the Internet through their enterprise LAN 94 and the LAN network
22 server 92.

23 It is noted that both implementations of Figs. 2 and 3 are shown and
24 described as suitable examples for implementing various aspects of the invention.
25 However, the network system might be implemented in a variety of arrangements.

1 In addition, the illustrations show the subscriber units as being personal computers
2 or work stations. However, the subscriber units can be implemented in other
3 forms which are capable of rendering content received over the network. As
4 examples, the subscriber computing units might include televisions, computers,
5 game devices, handheld devices, and the like.

6 As explained in the Background section, conventional techniques for
7 delivering video and audio content over the Internet is plagued with latency
8 problems. An aspect of this invention is to provide an improved method for
9 delivering streaming audio and video content over a network system. The
10 technique involves an intelligent, pre-caching and pre-loading of certain content at
11 the local service provider (e.g., ISP, POP, LAN network server) prior to optimal or
12 peak demand times when the content is likely to be requested by the subscribers.
13 In this manner, the frequently requested content is already downloaded and ready
14 for access from the subscribers before they actually request it. When it is finally
15 requested, the data can be streamed continuously in real-time for just-in-time
16 rendering from the local service provider to the subscriber. This eliminates the
17 latency problems of prior art systems. Moreover, intelligently pre-caching content
18 before peak demand times is more effective than traditional on-demand caching
19 because the content is available to the first subscriber who requests it.

20 Fig. 4 shows a functional block diagram of a local service provider 110
21 according to one implementation which enables intelligent pre-caching and pre-
22 loading. At its most fundamental level, the local service provider 110 provides an
23 on-ramp connection to the Internet for its subscribers. The subscribers send
24 requests to the local service provider 110 for content available on the Internet. The
25 local service provider acts as an intermediary facilitator which communicates the

1 requests to the appropriate content server and then returns the requested content to
2 the appropriate subscribers.

3 The local service provider 110 has a request handler 111 which manages
4 requests received from the subscribers. In the Web context, the subscriber
5 computers run Web browser applications which generate requests in the form of
6 universal resource locators (URLs). A URL describes everything about a
7 particular resource that a Web browser needs to know to request and render it.
8 The URL describes the protocol a browser should use to retrieve the resource, the
9 name of the computer it is on, and the path and file name of the resource. The
10 following is an example of a URL:

11
12 <http://www.microsoft.com/upgrades>
13

14 The “http://” portion of the URL describes the protocol. The letters “http”
15 stand for HyperText Transfer Protocol, the set of rules that a browser will follow
16 to request a document and the remote server will follow to supply the document.
17 The “www.microsoft.com” portion of the URL is the name of the remote host
18 computer which maintains the document. The last portion “/upgrades” is the path
19 and file name of the document on the remote host computer.

20 When the request handler 111 receives a request, the local service provider
21 110 first looks to its own cache memory 124 to determine if a proxy copy of the
22 target resource referenced by the URL is stored locally. The cache memory 124
23 serves as a quasi-temporary local storage for holding proxy copies of often used
24 and requested target resources. The cache memory 124 can be implemented using
25 different types of memory, including RAM, storage disks (optical, magnetic, etc.),

1 and the like. If a proxy copy is stored in the cache memory 124, the target
2 resource is served locally from the cache memory 124. If there is no proxy copy,
3 the local service provider 110 uses the URL request to locate the target resource
4 from a content provider and to request delivery of the target resource over the
5 Internet. The local service provider 110 passes the target resources on to the
6 requesting subscriber and may also cache the target resource in the cache 124 if
7 the policy rules governing the cache are met.

8 The local service provider 110 has a hit recorder 112 which is coupled to
9 receive the URLs submitted by the subscribers. For each URL, the hit recorder
10 112 records hit information in a URL hit database 114. The hit information
11 includes the date/time of the request, the subscriber who made the request, and
12 other information. The hit recorder 112 also triggers a pattern recognizer 116
13 which draws on information in the URL hit database 114 to detect repetitive access
14 behavior patterns based on subscriber requests. The pattern recognizer 116
15 performs statistical analyses using hit data from the URL hit database to determine
16 usage patterns that help the local service provider be more responsive to the needs
17 of its clientele. For instance, in the preferred implementation, the pattern recognizer 116
18 determines which URLs, and hence which Internet resources, are
19 being requested most often and least often, and the time of day when the most
20 requests are received. The pattern recognizer 116 is also responsive to operator
21 input to allow adjustment or tuning by the operator for specialized analysis.

22 A scheduler 118 uses the pattern results generated by the pattern recognizer
23 116 to schedule requests for specific URLs of target resources on the Internet. The
24 requests are scheduled to be filled at pre-selected times prior to the peak times
25 when the highest number of users are most likely to request the content found at

1 the URLs. Administrative tools 120 permit the operator to configure various
2 operating parameters.

3 The pattern recognizer 116 and scheduler 118 cooperate to enable
4 intelligent pre-caching of frequently requested content. The operation of the local
5 service provider 110 to perform this intelligent pre-caching according to an aspect
6 of this invention is described in conjunction with reference to the flow diagram of
7 Fig. 5. The local service provider is programmed to perform the computer-
8 implemented steps of Fig. 5 to alleviate the problems of providing streaming video
9 and audio data over the Internet. The steps are presented in the illustrated order
10 for discussion purposes, but are not restricted to this sequence.

11 The pattern recognizer 116 monitors the patterns of the subscriber requests
12 to determine which content is most frequently requested and when (step 150 in
13 Fig. 5). From these patterns, the pattern recognizer 116 can identify peak times in
14 subscriber traffic and the relation of the peak times to specific requested content
15 (step 152). For instance, suppose that a high number of subscribers frequently
16 request the CNN Web page during the morning hours of 6:30 AM to 8:00 AM.
17 These requests translate into a high number of URL hits for the CNN Web page
18 which are recorded by hit recorder 112 in the URL hit database 114. The pattern
19 recognizer 116 recognizes this recurring pattern of requests for the CNN Web page
20 and identifies the peak time for this Web page to be between 6:30 AM and 8:00
21 AM.

22 Using the patterns identified by the pattern recognizer 116, the scheduler
23 118 schedules delivery of the content at a selected time prior to the peak time (step
24 154 in Fig. 5). In this example, the scheduler 118 might schedule delivery of the
25 CNN Web page at a time prior to 6:30 AM. For instance, the scheduler 118 might

1 schedule a request for the CNN Web page at 6:00 AM to provide sufficient time to
2 downloaded that page before the earliest subscribers are expected to begin asking
3 for it, yet not too early to ensure that the latest news is included.

4 At the scheduled time, a media loader 122 sends a request to the content
5 server on the Internet and receives the content from that content server (step 156 in
6 Fig. 5). The content is stored locally at the local service provider (step 158).
7 More particularly, the data comprising the target resource is stored as a proxy file
8 in the cache memory 124, and any continuous data content (e.g., audio or video
9 data) is stored in the continuous media server 126. In the Web context, the content
10 might be in the form of a Web page or other hypermedia document that has
11 hyperlinks to various data items, such as audio and/or video clips. The
12 hypermedia document itself is stored in the cache memory 124, while the audio
13 and video clips referenced by the hyperlinks are stored in the CMS 126. The
14 target specifications corresponding to the links in the cached hypermedia
15 document are modified to reference the audio and video files in the CMS 126, as
16 opposed to the files maintained at the Web site (step 160 in Fig. 5). As an
17 alternative to modifying the target specifications, a conversion table can be
18 constructed which converts requests from referencing the files at the Web site to
19 referencing the files in the CMS 126.

20 In our CNN example, the local service provider 110 sends a request to the
21 CNN Web site seeking to download the CNN Web page at 6:00 AM. The CNN
22 Web page is downloaded over the Internet and stored in the cache memory 124. If
23 the CNN Web page contains links to any audio or video clips of recent news, these
24 data files are also downloaded and stored in the CMS 126. The links within the
25

1 cached Web page are modified to reference the audio and video files stored locally
2 in the CMS 126, instead of the files maintained at the CNN Web site.

3 The media loader 122 loads the locally stored content just before the peak
4 time so that it is ready to serve during the peak time (step 162 in Fig. 5). When the
5 first subscriber requests the CNN Web page at, for example, 6:40 AM, the local
6 service provider 110 serves the Web page from the cached memory 124. If the
7 subscriber activates a link to a video or audio file, the local service provider 110
8 immediately serves the data stream from the CMS 126 for just-in-time rendering
9 on the subscriber's computer. Accordingly, the video file is served as streaming
10 data to even the first subscriber who requests it, rather than making that subscriber
11 wait for the file to be retrieved over the Internet.

12 The intelligent pre-caching method obviates the latency problems
13 associated with streaming video and audio over the Internet, and is a further
14 improvement to traditional on-demand caching techniques. However, it is noted
15 that the network system does not accommodate data streaming for every video and
16 audio file on the Internet, but instead only selected files. The system makes an
17 intelligent choice as to which content is likely to be requested by its subscribers
18 and then makes only this content readily available to the subscribers. In this way,
19 the method seeks to optimize the physical computing resources of the local service
20 provider in a manner which best services the majority of the clientele.

21 It is noted that the content servers serve many local service providers over
22 the Internet. These local service providers, in turn, serve many different users.
23 Due to varying demographics, the local service providers will generally differ in
24 the content that it most often serves to its clientele. For example, a service
25 provider in Seattle, Washington, might have many requests for content on

1 entertainment or news local to Seattle. The pattern recognizer for a Seattle-based
2 service provider might therefore schedule proportionally more Seattle related
3 content than, say, a London-based service provider. As a result, the sets of pre-
4 cached content may differ significantly from one service provider to another
5 depending upon the results of the local hit recorder and pattern recognizer.

6 The system and method described above places the authority for deciding
7 which content is pre-cached at the local service provider. This allows the local
8 service provider to adapt to the often changing patterns of its clientele. However,
9 in another implementation, the content servers can be given the governing
10 authority of deciding when and what content to download to the ISPs prior to peak
11 times. For instance, the content server can maintain a schedule of when to
12 download different sets of content to various ISPs in timely fashion before the sets
13 of content are requested by the respective users who are serviced by the ISPs.

14 With continuing reference to Fig. 4, in this implementation, the local
15 service provider 110 also includes a policy manager 128 which defines and
16 administers rules that determine which documents or resources are cached in the
17 cache memory 124. For instance, caching rules might call for caching resources
18 that are routinely requested by many subscribers, but foregoing caching resources
19 that are rarely or infrequently requested. The policy rules also coordinate cache
20 maintenance by deciding when documents are out-of-date and how these
21 documents are deleted from the cache memory 124.

22 According to another aspect of this invention, time-to-live (TTL) tags are
23 assigned to the content to assist in determining when the content should be
24 refreshed or disposed. The TTL tags can be assigned by the content server as part
25

1 of the content itself. The server can attach an expiration tag which represents the
2 publisher's best estimate as to how long it will be before the content is updated.

3 Alternatively, the local service providers might compute the TTL tags for
4 the content it caches in cache memory 124. The computation is based upon a
5 theory that older content is less likely to change. Content that changed only 10
6 minutes ago is statistically more likely to change within the next 24 hours than
7 content that last changed one month ago. In one implementation, an approximate
8 TTL is computed as a percentage of time since the content is known to have last
9 changed. The percentage is an operator controlled parameter. Suppose a 10%
10 value is selected. Content that last changed 72 hours ago is assigned a TTL tag of
11 7.2 hours. If the content is not updated within 7.2 hours, it is given a new value of
12 7.9 hours (i.e., $(72 \text{ hours} + 7.2 \text{ hours}) \times 10\% = 7.9 \text{ hours}$). As the content ages, it
13 is checked less often. The TTL tags can be kept in a separate table of the cache
14 124 to correlate the tags and their content.

15 Deletion policies are a function of the content itself (e.g., its TTL tags), the
16 subscriber patterns (e.g., how frequently the content is requested), the cost to
17 request newer updated content, and the constraints imposed by capacity limitations
18 of the cache memory.

19 The local service provider 110 also maintains a subscriber database 130
20 which stores lists of subscribers (or LAN users in the LAN configuration) and
21 pertinent information about them (e.g., routing addresses, billing addresses, etc.).
22 A usage reporter 132 uses the URL hit information from the URL hit database 114
23 and subscriber information from the subscriber database 130 to generate reports on
24 subscriber usage patterns. These reports can be used by the operator to efficiently
25 allocate computer resources to best satisfy the needs of its clientele. The reports

1 can also be used by content providers to help them assess the popularity of their
2 Web sites and the type of subscribers who visit them.

3 In a preferred implementation, the functional components described with
4 respect to Fig. 4 are implemented in software which executes on the host computer
5 of the local service provider. It is noted that the functional layout is provided for
6 explanation purposes. The subscriber database 130, the URL hit database 114, and
7 the cache memory 124 can be implemented as one database server. Other
8 implementation variations may also be made.

9 In the above system, the local service providers (e.g., ISPs, LAN Web
10 servers) initiate the requests for content so that it may be pre-cached prior to peak
11 demand times when the content is most likely to be requested. This system can be
12 referred to as a "pull-caching" system in that content is pulled over the Internet
13 upon request of the local service providers. The method of intelligently pull-
14 caching data prior to peak times enables delivery of streaming video and audio
15 data to Internet users.

16 Fig. 6 shows a network system 200 according to another aspect of this
17 invention. The network system 200 attacks the latency problem of streaming video
18 and audio data by supplementing the primary Internet distribution network with a
19 second network which is not reliant on the Internet/ISP connection. The
20 Internet/ISP connection is often the bottleneck for streaming data and is typically
21 the connection least likely to be upgraded due to economic factors surrounding the
22 business of the ISP. Although not required, in this implementation, the content
23 may be pushed top down from the content provider over the Internet and thus, the
24 system may be referred to as a "push-caching" system.

1 Network system 200 is similar to the configuration of the Fig. 2 network
2 system 50 in that it has a content server 52 which serves content over a high-speed,
3 high-bandwidth network 54, via local ISPs 56, to end users 58 and 60. The
4 difference between the two systems is that network system 200 of Fig. 6 has an
5 additional, secondary network 202 for distributing content from the content server
6 52 to the ISPs 56. In the illustrated implementation, the secondary network 202 is
7 a broadcast satellite network. The content provider 52 has a transmitter 204 which
8 sends signals to an orbiting satellite 206, which redirects the signals to an ISP-
9 based receiver 208.

10 The secondary satellite network 202 affords a supplemental bandwidth for
11 delivery of content to participating ISPs in addition to the content delivered over
12 the interactive network connection 62. For instance, using present DSS (digital
13 satellite service) technology, the satellite network 202 provides an additional 6
14 Mbps bandwidth to deliver content to the ISP 56. This extra bandwidth is made
15 available at a fraction of the cost of buying T1/T3 connections.

16 The supplemental-caching technique allows the content provider to
17 download more information in a timely manner. To continue the above CNN
18 example, the CNN content provider might transmit the CNN Web page over the
19 satellite system 202 minutes before the peak time of 6:30 AM to 8:00 AM. The
20 ISP receives the Web page from its satellite receiver 208 and caches the Web page
21 for serving during the peak time. The CNN Web page is thereby efficiently made
22 available for real-time streaming to the subscribers, without tying up or consuming
23 any of the bandwidth provided by the network connection 62.

24 In a preferred implementation, the supplemental satellite network 202 is
25 uni-directional, in that data is broadcast from the content provider 52 to the ISP

1 56. It is a low cost solution increasing the bandwidth of the pipeline to the ISP,
2 without requiring significant investment on the part of the ISP. In addition to
3 satellite technologies, the broadcast network 202 can be implemented as other
4 wireless systems, such as RF or cellular technologies.

5 In another embodiment, the secondary network 202 can be implemented as
6 an second data communications network whereby supplemental content is
7 multicasted to participating ISPs prior to the peak time.

8 In compliance with the patent statutes, the invention has been described in
9 language more or less specific as to structure and method features. It is to be
10 understood, however, that the invention is not limited to the specific features
11 described, since the means herein disclosed comprise exemplary forms of putting
12 the invention into effect. The invention is, therefore, claimed in any of its forms
13 or modifications within the proper scope of the appended claims appropriately
14 interpreted in accordance with the doctrine of equivalents and other applicable
15 judicial doctrines.

16

17

18

19

20

21

22

23

24

25